Reporting: wrangle_report¶

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

Gathering Data¶

The primary data source was a Twitter account that rates dogs with humorous comments. The initial dataset contained fields like tweet IDs, timestamps, source of the tweet, text of the tweet, dog names, and some categorical variables indicating a dog's "stage" such as 'doggo', 'puppo', 'floofer', etc.

Twitter Archive File: This CSV file provided foundational data such as tweet IDs, text, date of posting, and more. Image Predictions File: Using the tweet IDs from the above file, we could access more information about the tweets, mainly the breed of the dog featured in each tweet. Twitter API Data: We used the tweet IDs to gather further data using the Twitter API, like the number of retweets and favorites.

Assessing Data¶

Upon an initial assessment, the following issues were identified:

Quality Issues:¶

1. Original Ratings with Images: We only want original ratings (no retweets) that have images.
2. Timestamp Clean-up: timestamp column, change dtype to to_date and delete +0000 in the end.
3. Name Anomalies: Some names are not mentioned in a tweet, so the values should be 'None', but in our case we still have some idefinite articles or different words that where picked right after "This is" beginig of the tweet. All the names starting with capital letter, all the rest garbage data starts with lowercase words. We'll replace lowercase words with 'None'.
4. Text Column Clean-up: text column, in the end of the text every post has a broken link, because of quotation symbol. Remove quotation symbol.

5. ID Datatype Adjustments: in_reply_to_status_id column, dtype change to integer.
6. ID Datatype Adjustments: in_reply_to_user_id column, dtype change to integer.
7. Rating Numerator: rating_numerator column, some values > 14, since 14 is the max rating given by author we should fix it.
8. Source Column Refinement: source column, get information inside the tags and remove tags.

Structural Issues:¶

1. Dog Stage Consolidation: column names 'doggo' , 'floofer', 'pupper', 'puppo' should be values for one column ex. "dog_stage"
2. Text-Link Separation: link in the end of 'text' column values should be in separate column.

Cleaning Data¶

Armed with our findings from the assessment phase:

Filtering Rows: We kept only rows that had images and were original ratings.

Refining Timestamp: The datatype was altered, and the redundant '+0000' was removed.

Name Rectification: Lowercase words mistaken as names were replaced with ‘None'.

Text Clean-up: The unwanted quotation symbols at the end of the text were removed.

Datatype Changes: Relevant ID columns were adjusted to integer datatypes.

Rating Corrections: Rating numerators exceeding the permissible value were rectified.

Source Clean-up: HTML tags from the source column were stripped, revealing just the source text.

Data Reshaping: Using the melt function from pandas, the dataset was reshaped to consolidate the dog stages. Additionally, the link from the text column was split into its dedicated column.

Conclusion¶

Through this data wrangling exercise, the significance of meticulous assessment and cleaning becomes evident. Such a process ensures that datasets are reliable and structured appropriately for analysis. The clean dataset, now saved as 'twitter_archive_master.csv', is ready for further exploration and analysis.